

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号
特開2003-345828
(P2003-345828A)

(43)公開日 平成15年12月5日(2003.12.5)

(51)Int.Cl. ⁷	識別記号	F I	テーマコード*(参考)
G 0 6 F 17/30	3 5 0	G 0 6 F 17/30	3 5 0 C 5 B 0 7 5
	1 7 0		1 7 0 A 5 B 0 9 1
	3 3 0		3 3 0 C
17/28		17/28	C

審査請求 未請求 請求項の数4 O L (全 8 頁)

(21)出願番号 特願2002-150721(P2002-150721)

(22)出願日 平成14年5月24日(2002.5.24)

(71)出願人 000004226

日本電信電話株式会社

東京都千代田区大手町二丁目3番1号

(72)発明者 笠原 要

東京都千代田区大手町二丁目3番1号 日

本電信電話株式会社内

(74)代理人 100066153

弁理士 草野 卓 (外1名)

Fターム(参考) 5B075 ND03 NK32 NK35 PP24 PQ36

PR06 QM08

5B091 AA15 CC02 CC04 CC15 CC16

(54)【発明の名称】 単語類似度計算方法、この方法を実施する装置、単語類似度計算プログラム、このプログラムを記録した記録媒体

(57)【要約】

【課題】 2つの言語内の任意の単語間の類似性判別を行う。

【解決手段】 一方の言語で使用される単語の複数それぞれについて座標を付与したデータより成る概念ベース11と他方の言語で使用される単語の見出し語を一方の言語で使用される単語の説明語で説明するデータより成る2言語辞典12を準備し、他言語概念ベース作成部13は2言語辞典12および概念ベース11を参照して見出し語の座標を決定してこれらを他言語概念ベース14に収納し、2言語概念ベース作成部15は概念ベース11および他言語概念ベース14を参照して含まれる他方の言語の見出し語に対する座標を取得して両者より全ての単語と対応する座標を2言語概念ベース16に収納し、類似度計算部17は2単語を外部より受け取り類似度を計算出力する単語類似度計算方法、この方法を実施する装置、単語類似度計算プログラム、このプログラムを記録した記録媒体。

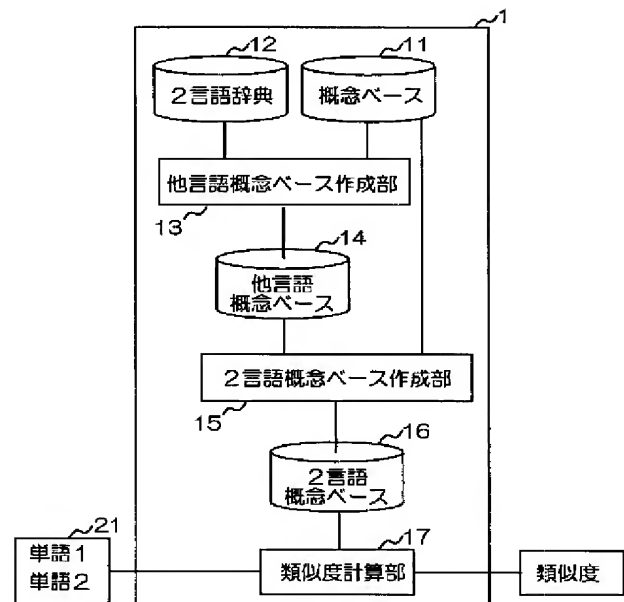


図1

【特許請求の範囲】

【請求項1】 2つの自然言語中の任意の単語2つが入力として外部より与えられた時に2単語間の類似度を計算する単語類似度計算装置において、一方の言語で使用される単語複数それぞれについて、多次元空間中での位置を表す座標が付与されているデータベースである概念ベースと、他方の言語で使用される単語より成る見出し語を一方の言語で説明する単語より成る説明語のデータベースである2言語辞典と、2言語辞典中の個々の見出し語に対応する説明語個々に対して、概念ベース中の説明語の座標を取得して、その結果に基づいて他方の言語の見出し語の座標を決定する他言語概念ベース作成部と、他言語概念ベース作成部によって作成された他方の言語の多次元空間の座標を表すデータベースである他言語概念ベースと、概念ベースと他言語概念ベースより、一方の言語と他方の言語の単語個々について、多次元空間中の座標を対応づけるデータベースである2言語概念ベースを作成する2言語概念ベース作成部と、概念ベースと他言語概念ベースより、一方の言語と他方の言語の単語個々について多次元空間中の座標を対応づけるデータベースである2言語概念ベースと、外部より入力された2単語に対して、各単語の多次元空間中の座標を取得して類似度を計算し、外部に出力する類似度計算部とを具備することを特徴とする単語類似度計算装置。

【請求項2】 2つの言語中の任意の単語2つが入力として外部より与えられた時に2単語間の類似度を計算する単語類似度計算方法において、一方の言語で使用される単語の複数それぞれについて座標を付与したデータより成る概念ベースと他方の言語で使用される単語の見出し語を一方の言語で使用される単語の説明語で説明するデータより成る2言語辞典を準備し、他言語概念ベース作成部は、2言語辞典を参照して、見出し語のそれぞれに対応する説明語のリストを取得し、各説明語のリストに対して概念ベースを参照して各説明語の座標を取得し、更に、各説明語に対して獲得された1或いは複数の座標より見出し語の座標を決定し、他言語概念ベース作成部より送られた他方の言語の単語の見出し語と座標を他言語概念ベースに収納し、2言語概念ベース作成部は、概念ベースを参照して含まれる一方の言語の単語全てに対する座標を取得すると共に他言語概念ベースを参照して含まれる他方の言語の見出し語に対する座標を取得し、他言語概念ベースおよび概念ベースより全ての単語と対応する座標を2言語概念ベースに収納し、ここで、類似度計算部は2単語を外部より受け取り、2

言語概念ベースを参照してそれぞれの単語の座標を取得し、これら2つの座標に基づいて類似度を計算し、同様にして、一方の単語に対して2言語概念ベース中の更なる他の単語の類似度を計算し、類似度を決定することを特徴とした単語類似度計算方法。

【請求項3】 他言語概念ベース作成部が、他方の言語で使用される単語の見出し語を一方の言語で使用される単語の説明語で説明するデータより成る2言語辞典を参照して見出し語のそれぞれに対応する説明語のリストを取得し、各説明語のリストに対して一方の言語で使用される単語の複数それぞれについて座標を付与したデータより成る概念ベースを参照して各説明語の座標を取得し、更に各説明語に対して獲得された1或いは複数の座標より見出し語の座標を決定し、他言語概念ベース作成部より送られた他方の言語の単語の見出し語と座標を他言語概念ベースに収納し、2言語概念ベース作成部は、概念ベースを参照して含まれる一方の言語の単語全てに対する座標を取得すると共に他言語概念ベースを参照して含まれる他方の言語の見出し語に対する座標を取得し、他言語概念ベースおよび概念ベースより全ての単語と対応する座標を2言語概念ベースに収納し、ここで、類似度計算部は、2単語を外部より受け取り、2言語概念ベースを参照してそれぞれの単語の座標を取得し、この2つの座標に基づいて類似度を計算し、一方の単語に対して2言語概念ベース中の更なる他の単語の類似度を計算して類似度を決定する指令を単語類似度計算装置の電子計算機に対して実行する単語類似度計算プログラム。

【請求項4】 請求項3に記載される単語類似度計算プログラムを記録した記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】この発明は、単語類似度計算方法、この方法を実施する装置、単語類似度計算プログラム、このプログラムを記録した記録媒体に関し、特に、人間が単語間の近さを判定する類似性判別の工学的なシミュレーションを実施する単語類似度計算方法、この方法を実施する装置、単語類似度計算プログラム、このプログラムを記録した記録媒体に関する。

【0002】

【従来の技術】従来例を図を参照して説明する。人間は、単語が表す意味を考慮して、指定された単語に対して複数の単語の中から類似した単語を選択したり、指定された単語に対して類似した単語を回答すること（以降、「類似性判別」と呼ぶ）ができる。数万以上の多数の単語を対象として、人間が行うこのような類似性判別をコンピュータを使用して再現する技術は、テキストに関する情報処理の基盤となる重要な技術の1つである。類似性判別技術およびそれに基づく単語のデータベース

は、コンピュータ、ネットワーク中に蓄積された膨大な規模のテキストデータを処理したり、人間の知的活動の一部を肩代りするために幅広く使用されている。例えば、情報検索、機械翻訳、オントロジー構築、ナレッジマネジメント技術で具体的に応用されている。

【0003】類似性判別の再現を実現する技術の1つとして、単語を多次元空間中に配置して、2つの単語の類似性の度合(以降、「類似度」と呼ぶ)を空間中の単語同志の近さの度合に基づいて計算する方法(以降、「ベクトルモデル」と呼ぶ)がある[Deerwester90, Schutze92, 笠原 97]。単語間の類似度としては、一般的に、空間中の単語の位置ベクトルのなす角度の余弦を類似度とする方法が採用されている。また、多次元空間の次元数は数百から数千に渡っている。この様な多次元空間中に多数の単語を人手で配置することは技術的に困難であるので、既存のテキストデータを使用して配置する技術が検討されている。例えば、[笠原 97]においては、国語辞典の見出し語の空間配置を説明文中の単語の出現頻度に基づいて決定している。また、[Deerwester 90]においては、単語が出現する検索文書の傾向に基づいて決定し、[Schutze 92]においては、文書中の2つの単語が同時に現れる傾向に基づいて決定している。これらは文書集合を対象として質問に適合する文書を探して出力する情報検索[Deerwester 90, Schutze 95, 熊本 99]、文書を内容に基づいて分割するテキストセグメンテーション[別所 01]に適用されている。

【0004】ベクトルモデルを使用して日本語と英語の様な複数の言語について、含まれる単語群を多次元空間中に配置して、空間中の任意の単語の類似度を計算することができれば、翻訳にかかわる様々な技術への利用が可能である。例えば、コンピュータを使用して或る言語の文を別の言語の文に変換する技術である機械翻訳においては、双方の言語で同じ意味を表す単語の対応関係より成る対訳辞書を技術の基盤としている。上記2言語の単語に関する多次元空間を使用して、1方の言語の単語に対して類似している他方の言語の単語を検索することができれば、上記対訳辞書を自動生成することが可能である。また、複数の言語で記述された文書群から情報を検索する多言語情報検索においては、様々な言語で記述された個々の文書、質問について、構成する単語の重心として多次元空間に配置することによって検索することが可能である。

【0005】

【発明が解決しようとする課題】上述した2つの言語それぞれを構成する単語の任意の組み合わせについて類似性判別を再現するに、ベクトルモデルを使用する場合は、2つの言語の単語が1つの多次元空間中に適正に配置されていることが要求される。[Schutze 92]の方法論を使用して2つの言語の単語を1つの多次元空間に配置する方法としては、文献[Masuichi 00]があげられる。

これは、一方の言語のテキストコーパスに対してもう一方の言語の対訳より成る2言語対訳コーパスを利用し、2言語の全単語間の共起関係の傾向に基づいて多次元空間中に配置する方法である。しかし、この技術で使用する2言語対訳コーパスは任意の2言語でもれなく十分に用意されていないので、技術の適用範囲は狭まることになる。

【0006】一方、辞典は、英和辞典、独和辞典、西和辞典の様に多種の2言語辞典が存在するので、これらを使用してベクトルモデルを適用することができれば、数多くの2言語の単語の多次元空間を作成することができ、2言語の類似性判別を可能とする。しかし、辞典を使用したベクトルモデルに関する技術を2言語の類似性判別の再現に使用するには、2つの言語の単語が見出し語として現われる辞典が必要である。実際は、その様な辞典は殆ど存在しないので、直接適用することはできない。また、容易に推測可能な適用方法としては、複数の辞典を組み合わせる上記の様な辞典を作成してモデルを適用する方法があろう。例えば、日本語と英語を対象とした場合、国語辞典と英和辞典とを用意して1つの辞典としてまとめる。このことによって、日本語の単語と英語の単語について、何れも日本語による説明文が付与された辞典とみなすことができるので、この辞典にベクトルモデルを適用することで、日本語と英語の単語を一緒に1つの多次元空間中に配置することができる。

【0007】しかし、一方の言語の単語を他方の言語の単語で説明する2言語辞典の説明文の傾向は、1つの言語の単語を同じ言語の単語で説明する国語辞典の説明文と傾向が異なるところから、上記の様な単純な適用は困難である。例えば、英和辞典は、見出し語'cow'に対して、「乳牛、牝牛」の様に対応する訳語のみが簡潔に記載されている。一方、国語辞典は、見出し語'乳牛'に対して、「主として牛乳を取るために家畜として飼育される牛の名称。……」の様に定義的な長文の説明文が対応しており、両者の辞典の説明文の長さは大きく異なっている。従って、説明文中の単語同志の出現傾向を比較することを基本とするベクトルモデルは、英和辞典に起因する'cow'と国語辞典に起因する'乳牛'の様な、意味を人間が考慮した場合類似していると考えられる異なる言語での単語同志は、単純に2つの辞書を足し合わせて作成した多次元空間中においては、互いに近接して配置はされない点が問題である。辞典とは、作成する目的によって記述方法、記述の長さが異なっているために、個々の辞典から作成される多次元空間同志を直接比較することは困難である。

【0008】以上を要約するに、従来は1種類の言語に関するテキストデータを使用してそれに含まれる単語間の類似度を計算していたベクトルモデルを使用して、2つの言語内の任意の単語間の類似性判別を行うことを考える場合、種類、形式、或いは規模の異なる2つのテキ

ストデータを単純に1つのテキストデータとしてまとめてベクトルモデルを適用することはできないので、如何にして2種類のテキストデータを併用するかが問題となる。この発明は、上述の問題を解消した単語類似度計算方法、この方法を実施する装置、単語類似度計算プログラム、このプログラムを記録した記録媒体を提供するものである。

【0009】

【課題を解決するための手段】この発明は、2つの自然言語（言語A、言語Bと呼ぶ）中の任意の単語2つが入力として外部より与えられた時に、2単語間の似ている度合を表す尺度である類似度を計算する単語類似度計算装置であり、言語Aで使用される単語を多次元空間中に配置し、また、言語Bで使用される単語多次元空間に配置する知識源とするために、言語Aで使用される単語複数それぞれについて、多次元空間中での位置を表す座標が付与されているデータベース（概念ベース）を備える。そして、言語Bで使用される単語を言語Aで用いられている単語と関連づけるために、言語Bで使用される単語（見出し語と呼ぶ）を言語Aで使用される単語（説明語と呼ぶ）で説明するデータベース（2言語辞典と呼ぶ）を備える。

【0010】また、言語Bで使用される単語を言語Aの単語に関する多次元空間中に配置する操作を行うために、2言語辞典中の個々の見出し語に対応する説明語個々に対して、概念ベース中の説明語の座標を取得して、その結果に基づいて言語Bの見出し語の座標を決定する部（他言語概念ベース作成部）を備える。更に、言語Bで使用される単語の多次元空間中の座標を参照するために、該他言語概念ベース作成部によって作成された言語Bの多次元空間中の座標を表すデータベース（他言語概念ベース）を備える。そして、A、B2つの言語の単語の多次元空間中の座標を一括して参照するための操作を行うために、該概念ベースと該他言語概念ベースより、言語Aと言語Bの単語個々について、多次元空間中の座標を対応づけるデータベース（2言語概念ベース）を作成する部を備える。

【0011】また、A、B2つの言語の単語の多次元空間中の座標を一括して参照するために、2言語概念ベースを備え、言語A、Bで使用される単語のいかなる組み合わせに対しても単語の類似度を計算するために、外部より入力されて言語A、或いは言語Bの単語1つのない複数の単語で構成された単語の集合2つ（単語1および単語2）に対して、単語1と単語2の多次元空間中の座標を計算し、これらの似ている度合である類似度を計算し、外部に出力する類似度計算部を備えている。 *

$$\text{Word}_i = (v_{i1}, v_{i2}, \dots, v_{in}) \dots\dots\dots (1)$$

属性として、概念ベース中の単語すべて（n語）を用いており、概念ベース全体は、属性の重みを要素とするn行n列の行列（G₁）となる。また、n語の属性をm個※50

*【0012】

【発明の実施の形態】この発明の実施の形態を図1の実施例を参照して説明する。この発明による単語類似度計算装置は2つの任意の自然言語である一方の言語Aと他方の言語Bを構成する任意の2つの単語を入力として、2つの単語の類似の度合を表す数値である類似度を出力する。従来の単語類似度計算装置1は個別の言語A或いは言語Bを内の単語同志の類似性判別のみを行うことができるに過ぎないが、この発明による単語類似度計算装置は、類似度を計算する対象である2つの単語は一方の言語Aの2単語、他方の言語Bの2単語とするのみならず、一方の言語Aと他方の言語Bに亘って言語Aの単語と言語Bの単語の2単語を組み合わせて類似度を計算する対象とすることができる。そして、この単語類似度計算装置は電子計算機を主要な構成要素として構成されている。

【0013】この単語類似度計算装置1は、一方の言語Aで使用される単語複数それぞれについて、多次元空間中における位置を表す座標が付与されているデータベース（概念ベース）11と、他方の言語Bで使用される単語（見出し語）を言語Aで使用される単語（説明語）で説明するデータベース（2言語辞典）12と、2言語辞典12中の個々の見出し語に対応する説明語個々に対して、概念ベース11中の説明語の座標を取得して、その結果に基づいて言語Bの見出し語の座標を決定する他言語概念ベース作成部13と、他言語概念ベース作成部13によって作成された言語Bの多次元空間における座標を表すデータベースである他言語概念ベース14と、概念ベース11と他言語概念ベース14より、言語Aと言語Bの単語個々について、多次元空間中の座標を対応づけるデータベースである2言語概念ベースを作成する部である2言語概念ベース作成部15と、概念ベース11と他言語概念ベース14より、言語Aと言語Bの単語個々について、多次元空間中の座標を対応づけるデータベースである2言語概念ベース16と、外部の単語入力源21より入力された言語Aと言語B中の2単語に対して、単語1、単語2の多次元空間中における座標を取得して、類似する度合である類似度を計算し、外部に出力する類似度計算部17とを具備している。

【0014】以下、単語類似度計算装置1およびその動作について説明する。まず、概念ベースについて説明する。概念ベースとは単語を属性の重みを表す実数を要素とするベクトル（「属性ベクトル」）で表現した知識ベースである。概念ベース中の単語W_i（i=1、…、n）の属性ベクトルW_{ordi}は以下の通りとなる。

※のカテゴリーに分離するシソーラス（類語辞典）を用い、同じ分類に含まれる属性をカテゴリーに一般化する。

【0015】

$$\text{Word}'_i = (v'_{i1}, \dots, v'_{ik}, \dots, v'_{im})$$

$$v'_{ik} = \sum v_{i1} T(1, k) \quad \text{但し, } 1 = 1 \sim n \dots \dots (2)$$

$T(1, k)$ はシソーラスを表す関数であり、1番目の属性がk番目のカテゴリーに含まれるときは1、それ以外は1を取る。 $(1, k)$ の要素の値を $T(1, k)$ としたn行m列の行列をTとすれば、シソーラスで属性を一般化した概念ベース全体は、n行M列の行列($G_2 = G_1 T$)となる。属性の重みは国語辞典の見出し語に対する説明文中の単語の出現頻度に基づいて獲得する。獲*10

$$\text{sim}(W_i, W_j) = \text{Word}'_i \cdot \text{Word}'_j = \sum v'_{ik} v'_{jk}$$

$$\text{但し, } k = 1 \sim m \dots \dots (3)$$

現在は、学研 国語大辞典[金田一 88]と30万語を3000カテゴリーに分類したシソーラス[池原 97]を用い、約9万語の概念ベースが自動構築されている[永森 00]。実施例における概念ベース11は、この一般的な概念ベースにおけるカテゴリーが2の場合に対応する。

【0016】概念ベース11は、言語Aに含まれる複数の単語それぞれについて、多次元空間中の座標が予め付与されたもののデータベースである。多次元空間とは次元数が1、2、3、或いはそれ以上の任意の次元数を持つ空間であり、個々の単語の多次元空間中の座標は関連する単語同志は互いに近接して設定されている。座標は、多次元空間中の次元数と同じ数の要素より成り、2次元であれば要素数は2である。概念ベース11に含まれる単語の数は、2単語以上ならば何単語であっても差し支えない。この実施例において使用する概念ベース11中の単語に付与される座標は人間が決定したものでも差し支えないし、文献[笠原 97]に記載される様な国語辞書より自動的に決定した座標であっても差し支えないし、文献[Schutze 92]の様な新聞記事その他のテキストコーパスを使用して自動的に決定したものであっても差し支えない。この概念ベース11は、言語Aの単語を受け取り、他言語概念ベース作成部13および2言語概念ベース作成部15に対して対応する座標を出力する。入力された単語が存在しないときは、原点の座標を出力する。

【0017】2言語辞典12は、言語Bで使用されている単語である見出し語に対して、言語Aで説明するために使用される単語(説明語)が列挙されたデータベースである。説明語としては、見出し語の訳語1語ないし複数語であっても差し支えないし、言語Aを日本語、言語Bを英語としたときの英和辞典中の英語の見出し語を日本語で説明した説明文を元として形態素解析を行い、名詞、動詞、形容詞の様な類似性判別に関わる単語を抽出した結果であっても差し支えない。他言語概念ベース作成部13は、先ず、2言語辞典12から言語Bの見出し語毎に、それに対する言語Aの説明語を読み取る。次に、概念ベース11を参照して言語Aの説明語各々の座※50

*得方式の詳細は文献[笠原 97]を参照されたい。なお、獲得された属性ベクトルそれぞれについて、個々の重みは正規化しておく($\sum v'_{ik} = 1$ 但し、 $k = 1 \sim n$)。これを用いて概念ベースに含まれる W_i, W_j ($1 \leq i, j \leq n$)の類似度 sim ($0 \leq \text{sim} \leq 1$)を対応する属性ベクトル Word'_i のなす角度の余弦で表す。

※標を読み込む。座標の記述されていない説明語の場合は、これ以降の見出し語の座標を決定する処理の対象外とする。全ての説明語について概念ベース11中に対応する座標が記述されていない場合、或いは、全ての説明語の座標が原点の座標の場合は、単語リストの座標を原点の座標とする。

【0018】原点以外に位置する1つ以上の単語の座標が得られた場合、その座標を平均して見出し語の座標とする。平均の座標を計算する方法として、個々の次元毎に座標の要素を加算平均して得られた座標を平均の座標とする方法、個々の単語の座標の要素に対して先ず要素の2乗和で除し、その結果を個々の次元毎に座標の要素を加算平均して得られた座標を平均の座標とする方法、その他の平均の座標が個々の単語の座標と等しく近くなる方法であれば何れであっても差し支えない。但し、複数の単語中に同じ座標を持つ単語が複数存在する場合は、それらの単語の座標は他の単語の座標よりも単語リストの座標に近くなる。上述の方法により取得された見出し語の座標を見出し語と対応づけて他言語概念ベース14に収録する。

【0019】他言語概念ベース14は、他言語概念ベース部13より与えられた、言語Bの複数の見出し語と各々に対応する座標より成るデータベースであり、2言語概念ベース作成部15に対して言語Bの見出し語に対応する座標を与える。2言語概念ベース作成部15は、概念ベース11を参照して含まれる言語Aの単語全てに対する座標を取得する。また、他言語概念ベース14を参照して含まれる言語Bの単語全てに対する座標を取得する。これら全てを2言語概念ベース16に出力する。2言語概念ベース16は、2言語概念ベース作成部15より出力された言語Aと言語Bの単語と対応する座標を受け取り収納する。類似度計算部17より言語A、或いは言語Bの単語を指定された時、対応する座標を検索して出力することができるならば、如何なる形式の収納であっても差し支えない。但し、含まれない単語を指定された時は、原点の座標を出力する。

【0020】類似度計算部17は、単語類似度計算装置1の外部の単語入力源21より入力される言語A、言語

Bの2つの単語、単語1および単語2を受け取る。次に、2つの単語のそれぞれに対して2言語概念ベース16を参照して2つの座標を取得する。そして、2つの座標に基づいて類似度を計算してこれを当該単語類似度計算装置1の外部に出力する。類似度の計算方法としては、2つの座標の同じ次元毎の要素の値の差の絶対値を加算した値の逆数（但し、同じ座標の場合は無限大とする）、2つの座標の同じ次元毎の要素の値の差の2乗和の逆数（但し、同じ座標の場合は無限大とする）、2つの座標の位置ベクトルのなす角度の余弦とする他の、座標同志が近接する程類似度の値が大きくなる計算方法であるならば、如何なる計算方法であっても差し支えない。

【0021】上述した単語類似度計算装置1およびその動作を具体例について更に具体的に説明する。言語Aは日本語、言語Bは英語とする。概念ベース11としては、図2に記載されるものを使用する。これは2次元平面に単語を配置したものである。また、2言語辞典12としては、英語の見出し語に対して訳語を列挙した図3に記載されるものを使用する。ここで、外部より英語、日本語の単語2語が与えられる先だって、概念ベース11と2言語辞典12を使用して2言語概念ベース16を予め作成しておく。先ず、他言語概念ベース作成部13は、2言語辞典12を参照して、見出し語'cow'、'bull'、'bird'のそれぞれに対応する説明語のリスト'牝牛'、'乳牛'、'雄牛'、および'鳥'を取得する。次に、個々の見出し語に対応する説明語のリストに対して、概念ベース11を参照して各説明語の座標を取得する。例えば'cow'の場合は、説明語'牝牛'の座標[0.6,0.7]と、'乳牛'の座標[0.55,0.6]を取得する。そして、見*30

$$\text{類似度} = (0.575 \times 0.55 + 0.65 \times 0.6) / \sqrt{(0.575^2 + 0.65^2)} \sqrt{(0.55^2 + 0.6^2)} \\ \approx 0.99988$$

この様に日本語の単語'牝牛'と英語の単語'cow'の類似度を計算することができる。

【0024】同様に、'cow'に対して、2言語概念ベース16中の単語'牛'、'乳牛'、'雄牛'、'鳥'、'bull'、'bird'の類似度を計算すると、'牛'=0.99813、'乳牛'=0.99984、'雄牛'=0.96676、'鳥'=0.82539、'bull'=0.96676、'bird'=0.82539となり、2言語概念ベース16中で'cow'に類似する日本語の単語を大きさの順に並べると、'牝牛'、'乳牛'、'牛'、'雄牛'、'鳥'と求めることができる。従って、この単語類似度計算装置1を他言語類似語検索に利用すれば、英語'cow'に対する日本語の類似語を'牝牛'と決定することができる。図3の2言語辞典を辞典と見なして文献[笠原 97]に記載される方法で概念ベースを作成すること自体は可能である。しかし、英和辞典から単純に作成された英語の概念ベースは、数語程度の説明語から作成された属性ベクトルより成る。一方、国語辞典から作成された概念ベースは、1つの単語について数十から数百の単語を説明語と※50

* 出し語に対して獲得された座標複数[0.6,0.7]、[0.55,0.6]より他言語概念ベース作成部13は、見出し語の座標を決定する。ここで、各座標の次元毎に値を平均したものを見出し語の座標とする。従って、 $[(0.6+0.55)/2, (0.7+0.6)/2] = [0.575, 0.650]$ が見出し語'cow'の座標となる。同様に、'bull'の座標は、概念ベース11の座標そのままの[0.90,0.60]、'bird'の座標も概念ベース11の座標そのままの[0.10,0.80]と決定される。決定した各々の座標と見出し語を他言語概念ベース14に送る。

【0022】他言語概念ベース14は、他言語概念ベース作成部13より送られた言語Bの単語の見出し語と座標を図4の様に収納する。2言語概念ベース作成部15は、概念ベース11を参照して含まれる言語Aの単語全てに対する座標を取得すると共に、他言語概念ベース14を参照して含まれる言語Bの単語全てに対する座標を取得する。これら全てを2言語概念ベース16に出力する。2言語概念ベース16は、他言語概念ベース14、概念ベース11より全ての単語と対応する座標を取得し、2言語概念ベース16に収納する。この様な2言語概念ベース16を図5に示す。

【0023】ここで、以上の単語類似度計算装置1に外部より単語'cow'、'牝牛'が与えられた場合について説明する。単語類似度計算装置1は、類似度計算部17を介してこの2単語を外部より受け取り、2言語概念ベース16を参照してそれぞれの単語の座標[0.575,0.65]と[0.55,0.6]を取得する。次に、この2つの座標に基づいて類似度を計算する。ここにおいては、2つの座標の位置ベクトルの余弦を類似度とする。従って、以下の様に求められる。

※して保有している国語辞典から作成される。従って、数語程度の説明語から決定される座標と多数の説明語から決定される座標の性質が異なることから、類似度計算の比較は困難である。

【0025】これに対して、この発明の単語類似度計算装置1の実施例は、2言語辞典だけを概念ベース作成にするのではなく概念ベースを参照して2言語辞典から英語の概念ベースを作成しているため、2言語辞典の説明語の多数に関わらない英語の座標を決定することができる。

【0026】

【発明の効果】以上の通りであって、この発明は、2つの自然言語、言語Aおよび言語Bの中の任意の単語2つが入力として外部より与えられた時に、2単語間の類似の度合を表す尺度である類似度を計算する単語類似度計算装置であり、従来の単語類似度計算装置の如く個別の言語内の単語同志の類似性判別を行うだけでなく、2つの言語AおよびBに亘って単語の任意の組み合わせで

単語の類似度を計算することができる単語類似度計算装置である。言語Aで使用される複数の単語それぞれについて、多次元空間中における位置を表す座標が付与されているデータベースである概念ベースを使用することにより、単語同志の類似の度合を数値で表現することができる。

【0027】言語Bで使用される単語である見出し語を言語Aで使用される単語である説明語で説明するデータベースである2言語辞典と、2言語辞典中の個々の見出し語に対応する説明語個々に対して、概念ベース中の説明語の座標を取得して、その結果に基づいて言語Bの見出し語の座標を決定する部である他言語概念ベース作成部を保有し、他言語概念ベース作成部によって作成された言語Bの多次元空間での座標を表すデータベースである他言語概念ベースを装置内部で作成するために、少ない説明語で記載されており、概念ベースを作成するには困難な2言語辞典からでも言語Aの単語と類似度を計算する言語Bの単語の座標を与えることができる。

【0028】そして、概念ベースと他言語概念ベースより、言語Aと言語Bの単語個々について、多次元空間中の座標を対応づけるデータベースである2言語概念ベースを作成する部である2言語概念ベース作成部と、概念ベースと他言語概念ベースより、言語Aと言語Bの単語個々について、多次元空間中の座標を対応づけるデータベースである2言語概念ベースを保有するので、2言語の任意の単語の座標を画一的に取得することができる。また、外部より入力された言語Aと言語B中の2単語に対して、単語1、単語2の多次元空間中での座標を取得して、似ている度合である類似度を計算し、外部に出力する類似度計算部を備えるので、2言語の任意の単語の類似度を計算することができ、情報検索等、2言語のテキストに関する情報処理への利用が容易となる。

【0029】【参考文献】・[Deerwester 90] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R.: Indexing by latent semantic analysis, Journal of the American Society for Information Science, Vol. 41, pp. 391-407 (1990).

・[Masuichi 00] Masuichi, H., Flounoy, R., Kaufmann, S., and Peters, S.: A Bootstrapping method for Extrac

ting Bilingual Text Pairs, in Coling 2000, pp. 1066-1070 (2000).

・[Schutze 92] Schutze, H.: Dimensions of Meaning, in Proceeding of Supercomputing 92, pp. 787-796 (1992).

・[Schutze 95] Schutze, H. and Pedersen, J.: Information retrieval based on word senses, in Fourth Annual Sympo. on Document Analysis and Information Retrieval, pp. 161-175 (1995).

10 ・[笠原 97] 笠原, 松澤, 石川: 国語辞書を利用した日常語の類似性判別, 情報処理学会論文誌, Vol. 38, No. 7, pp. 1272-1284 (1997).

・[熊本 99] 熊本, 島田, 加藤: 概念ベースの情報検索への適用ー概念ベースを用いた検索特性の評価ー, 情報研報, 第SIG-ICS115巻, pp. 9-16 (1999).

・[別所 01] 別所克人: 単語の概念ベクトルを用いたテキストセグメンテーション, 情報処理学会論文誌, Vol. 42, No. 11 (2001).

20 ・[永森 00] 永森, 笠原, 松澤: 概念ベース構築における表記と概念のマッピング手法, 人工知能学会全国大会, 第14巻, pp. 163-164 (2000).

・[金田一 88] 金田一, 池田 (編): 学研 国語大辞典 第二版, 学習研究社 (1988).

・[池原 97] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林 (編): 日本語彙大系, 岩波書店 (1997).

【図面の簡単な説明】

【図1】実施例を説明する図。

【図2】概念ベースを示す図。

30 【図3】2言語辞典を示す図。

【図4】他言語概念ベースを示す図。

【図5】2言語概念ベースを示す図。

【符号の説明】

11	概念ベース	12	2言語辞典
13	他言語概念ベース作成部	14	他言語概念ベース
15	2言語概念ベース作成部	16	2言語概念ベース
17	類似度計算部		

【図2】

単語	座標
牛	[0.7, 0.7]
牝牛	[0.6, 0.7]
乳牛	[0.55, 0.6]
雄牛	[0.9, 0.6]
鳥	[0.1, 0.8]

図2

【図3】

見出し語	説明語
cow	牝牛, 乳牛
bull	雄牛
bird	鳥

図3

【図1】

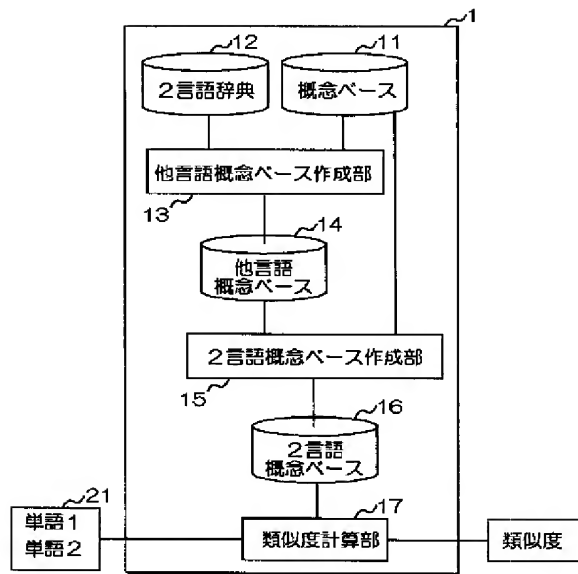


図1

【図4】

見出語	座標
cow	[0.575, 0.65]
bull	[0.9, 0.6]
bird	[0.1, 0.8]

図4

【図5】

単語	座標
牛	[0.7, 0.7]
牝牛	[0.6, 0.7]
乳牛	[0.55, 0.6]
雄牛	[0.9, 0.6]
鳥	[0.1, 0.8]
cow	[0.575, 0.65]
bull	[0.9, 0.6]
bird	[0.1, 0.8]

図5